University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2013 *Paper* 104

Covariance-Enhanced Discriminant Analysis

Peirong Xu* Ji Zhu†

Lixing Zhu‡ Yi Li**

*Southeast University

†University of Michigan - Ann Arbor

‡Hong Kong Baptist University

**University of Michigan - Ann Arbor, yili@umich.edu

Covariance-Enhanced Discriminant Analysis

Peirong Xu, Ji Zhu, Lixing Zhu, and Yi Li

Abstract

Linear discriminant analysis (LDA), a classical method in pattern recognition and machine learning, has been widely used to characterize or separate multiple classes via linear combinations of features. However, the high-dimensionality of the high-throughput features obtained from modern biological experiments, for example, microarray or proteomics, defies traditional discriminant analysis techniques. The possible interfeature correlations present additional challenges and are often under-utilized in modeling. In this paper, by incorporating the possible inter-feature correlations, we propose a Covariance-Enhanced Discriminant Analysis (CEDA) method that simultaneously and consistently selects informative features and identifies the corresponding discriminable classes. We show that, under mild regularity conditions, the proposed method can achieve consistency in parameter estimation as well as in model selection, and attain asymptotic optimal misclassification rate. Extensive simulations have verified the utility of the method. We have applied the method to study a renal transplantation trial, which was designed to identify genomic signatures that can identify kidneys with various functional types, a crucial step in drug development.

Covariance-Enhanced Discriminant Analysis

Peirong Xu[1], Ji Zhu[2], Lixing Zhu[3] and Yi Li[4] *

[1]Department of Mathematics, Southeast University, Nanjing, China

[2]Department of Statistics, University of Michigan, Ann Arbor, USA

[3]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

[4]Department of Biostatistics, University of Michigan, Ann Arbor, USA

Abstract: Linear discriminant analysis (LDA), a classical method in pattern recognition and machine learning, has been widely used to characterize or separate multiple classes via linear combinations of features. However, the high-dimensionality of the high-throughput features obtained from modern biological experiments, for example, microarray or proteomics, defies traditional discriminant analysis techniques. The possible inter-feature correlations present additional challenges and are often under-utilized in modeling. In this paper, by incorporating the possible inter-feature correlations, we propose a Covariance-Enhanced Discriminant Analysis (CEDA) method that simultaneously and

*The research described here was supported by a scholarship from the China Scholarship Council, a grant from the Research Grants Council of Hong Kong, a grant from the US NSF, and a grant from the US NCI.

consistently selects informative features and identifies the corresponding discriminable classes. We show that, under mild regularity conditions, the proposed method can achieve consistency in parameter estimation as well as in model selection, and attain asymptotic optimal misclassification rate. Extensive simulations have verified the utility of the method. We have applied the method to study a renal transplantation trial, which was designed to identify genomic signatures that can identify kidneys with various functional types, a crucial step in drug development.

Key words: linear discriminant analysis, pairwise fusion, correlation, graphical lasso, variable selection.

1 Introduction

Rapid advances in modern biological technology have yielded vast amount of high-throughput data, e.g. those arising from microarray or proteomics, which has brought a high demand in statistical methods that can effectively utilize such big data to make proper decision rules. For example, in a kidney transplantation and injury study (Flechner et al. 2004) that motivated this paper, 62 tissue samples were obtained from subjects with 4 different renal functional types after kidney transplantation. Distinguishing these 4 types of subjects based on 12,625 gene expression profiles is crucial to balance, at the molecular level, the need for immunosuppression to prevent transplanta-

tion rejection, while minimizing drug-induced toxicities. Linear discriminant analysis (LDA), a popular method in the classical setting where the number of variables is much smaller than the sample size, has been found to perform poorly in the high-dimensional setting because

(a) the sample covariance matrix, which is needed in LDA, is singular;

(b) the classification rule involves a linear combination of all the variables, causing difficulty in interpretation as well as degrading the classification performance with many non-informative variables.

As a remedy to address challenge (a), LDAs with a variety of penalized versions of covariance matrices have been constructed. They, for example, include the nearest shrunken centroids (NSC) method assuming the covariance matrix being diagonal (Tibshirani et al. 2002), the naive Bayes method using the diagonal of the sample covariance matrix (Bickel and Levina 2004), the extension of NSC with a general covariance matrix (Guo et al. 2007), the thresholding of mean effects and covariance matrix in binary classification (Shao et al. 2011), a Lasso-type classifier (Tibshirani 1996) based on the estimated product of mean effects and the precision matrix (Cai and Liu 2011). Other work, in similar contexts, include Qiao et al. (2008), Clemmensen et al. (2011), Witten and Tibshirani (2011) and the references therein.

On the other hand, to address challenge (b), Tibshirani et al. (2002) proposed the NSC method by shrinking the class centroids towards the global centroid, Wang and Zhu (2007) represented the NSC method as a Lasso

regression and introduced two new penalties to improve the effectiveness of variable selection, Guo (2010) used LDA with pairwise fusion penalties to select informative variables, while theoretical properties are in general elusive for these methods. Some asymptotic results are available for the annealed independence rule (FAIR) proposed by Fan and Fan (2008) and a LDA rule using penalized sparse least squares proposed by Mai et al. (2012). Note, however, both focus on binary classification and it is not clear how to extend them to the multiple class cases.

Overall, the above mentioned methods either use a diagonal matrix to approximate the covariance matrix, which ignores the correlation structure of the variables, or fail to perform variable selection, which may not be ideal for interpretation and classification accuracy.

In this paper, we propose a covariance-enhanced discriminant analysis (CEDA) method for high-dimensional multi-class classification. Our method utilizes the general covariance structure, going beyond the diagonal restriction, when selecting informative variables for LDA. We require the inverse of the covariance matrix, rather than the covariance itself, to be sparse. This is a much weaker assumption than those employed in previous work on LDA-based variable selection methods. Further, in terms of variable selection, we offer more flexibility by allowing a variable to be informative for only a subset of, rather than all, classes. Our work advances the field in several aspects. Firstly, it takes into account of the correlation structure between the variables and allows for simultaneously selecting informative variables

and identifying the corresponding discriminable classes. Secondly, we show the proposed procedure enjoys consistency of both parameter estimation and model selection. For binary classification, we also show that the proposed procedure achieves the asymptotic optimality in terms of the misclassification rate.

To further illustrate the impact of a non-diagonal covariance matrix, we consider a simple binary classification example as shown in Figure 1, wherein the two classes have the same mean in X_2 and different means in X_1. The best classifier would involve both X_1 and X_2 even though the latter by itself does not have any power in separating the two classes. Note that in this case, X_2 should still be considered as informative for classification and should not be removed by a variable selection method. The contribution of X_2 to classification is through its correlation with X_1, which demonstrates the role of using a non-diagonal covariance matrix in both classification and variable selection.

The rest of the paper is organized as follows. We introduce the framework and the proposed methodology in Section 2. We present the asymptotic properties in Section 3 and provide an algorithm for implementation in Section 4. We assess the finite sample performance of the proposed method and compare it with competing methods via simulation studies in Section 5. We apply the method to a data example arising from a kidney transplant rejection study in Section 6, and conclude the paper with Section 7. All technical proofs are relegated to the Supplemental Material.

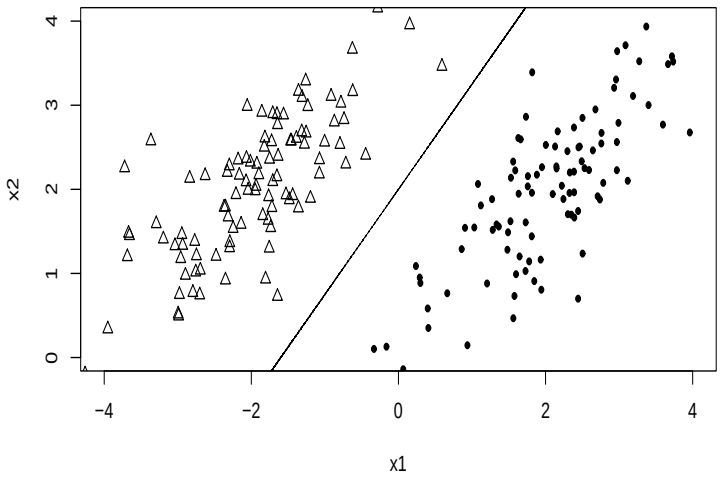

Figure 1: An illustrative example with two classes. Note that even though the two classes have the same mean in X_2, X_2 should still be considered as an informative variable for both classification and variable selection and should not be removed by a variable selection method.

2 Methodology

2.1 Model and notations

Consider a general K-class problem, where Y is the class label taking values in $\{1, 2, \ldots, K\}$ and X is the corresponding p_n-dimensional vector of predictors. We assume that the prior probability of class k is $\omega_k = P(Y = k) > 0$ for $1 \leq k \leq K$ satisfying $\sum_{k=1}^{K} \omega_k = 1$. The conditional density of X given class k is modeled by a multivariate Gaussian distribution, i.e. $X|Y = k \sim N_{p_n}(\mu_k, \Sigma)$, where $\mu_k = (\mu_{k1}, \ldots, \mu_{kp_n})^\tau$ is the class-specific mean vector and Σ is a $p_n \times p_n$ positive definite covariance matrix with (j, j')th element $\sigma_{jj'}$, $1 \leq j, j' \leq p_n$. As assumed in LDA, the covariance matrix Σ is a constant across different classes, which may be plausible as, for example, gene expressions across disease classes often differ in the means rather than in the covariance structure (Guo et al. 2010).

Let $\omega = (\omega_1, \ldots, \omega_K)^\tau$ and Ω be the inverse of Σ with (j, j')th element $\Omega_{jj'}$, $1 \leq j, j' \leq p_n$. Further, let $\mu = (\mu_1^\tau, \ldots, \mu_K^\tau)^\tau$ be the vector containing all class means and $x = (x_1, \ldots, x_{p_n})^\tau$ be an observation.

Given ω_k, μ_k for $1 \leq k \leq K$ and Ω (or Σ), the LDA is to classify an observation x to a class, say k^*, that maximizes

$$P(Y = k|X = x) = c(x)\omega_k \exp\left\{-\frac{1}{2}(x - \mu_k)^\tau \Omega (x - \mu_k)\right\},$$

where $c(x)$ is a normalizing constant that does not depend on k. For the purpose of variable selection, we compare two classes k and l, where $k \neq l$ with $k, l = 1, \ldots, K$. Specifically, we consider the pairwise difference for

$k \neq l$:

$$\begin{aligned} & \log P(Y=k|X=x) - \log P(Y=l|X=x) \\ = \; & (\log \omega_k - \log \omega_l) - \frac{1}{2}\sum_{j=1}^{p_n}\sum_{j'=1}^{p_n}\Omega_{jj'}(\mu_{kj}+\mu_{lj})(\mu_{kj'}-\mu_{lj'}) \\ & + \sum_{j=1}^{p_n} x_j \left\{ \sum_{j'=1}^{p_n} \Omega_{jj'}(\mu_{kj'}-\mu_{lj'}) \right\}. \end{aligned}$$

Note from the above equation that if variable j is non-informative for differentiating classes k and l, the necessary and sufficient condition is

$$\sum_{j'=1}^{p_n} \Omega_{jj'}(\mu_{kj'}-\mu_{lj'}) = 0. \tag{2.1}$$

Further we note that a sufficient condition leading to (2.1) is, for $j' = 1, \ldots, p_n$,

$$\begin{aligned} \Omega_{jj'} = 0 \text{ or } \mu_{kj'} - \mu_{lj'} = 0 \quad & \text{if } j' \neq j \\ \mu_{kj} - \mu_{lj} = 0 \quad & \text{if } j' = j \end{aligned} \tag{2.2}$$

Since $\Omega_{jj'} = 0$ indicates the conditional independence between X_j and $X_{j'}$ given all other variables, (2.2) implies that if a variable is conditionally independent ($\Omega_{jj'} = 0$) of all the variables differentiable for classes k and l ($\mu_{kj'} \neq \mu_{lj'}$), and is itself non-differentiable for classes k and l ($\mu_{kj} = \mu_{lj}$), it is then non-informative for differentiating classes k and l. This key observation motivates us to construct a variable selection procedure for selecting informative variables and identifying the discriminable classes simultaneously, which is presented in the next subsection.

2.2 Covariance-enhanced discriminant analysis (CEDA)

Let (y_i, x_i) be the ith observation $(i = 1, \ldots, n)$ from a K-class problem with known class label y_i and predictor vector x_i. Denote by $S(\mu) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i = k)(x_i - \mu_k)(x_i - \mu_k)^\tau$. A natural approach for inference is to maximize the log-likelihood function, which can be written as

$$l_n(\omega, \mu, \Omega) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i = k) \log \omega_k + \frac{1}{2} \log |\Omega| - \frac{1}{2} \mathrm{tr}(S(\mu)\Omega).$$

With high-dimensional parameters μ and Ω, a direct maximization is not stable. Regularization terms on μ and Ω are needed to enhance stability.

Motivated by condition (2.2), we propose to regularize the pairwise differences in class centroids for each variable and the off-diagonal elements of the concentration matrix. Specifically, we consider to maximize the following criterion

$$Q_n(\omega, \mu, \Omega) = l_n(\omega, \mu, \Omega) - \lambda_{1n} \sum_{j=1}^{p_n} \sum_{1 \le k < l \le K} |\mu_{kj} - \mu_{lj}| - \lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}| \quad (2.3)$$

$$\text{subject to} \qquad \sum_{k=1}^{K} \omega_k = 1 \text{ and } \Omega \succ 0 \quad (2.4)$$

where $\succ 0$ indicates positive definiteness. Note that the first penalty term in (2.3) shrinks the pairwise differences in class centroids for each variable, whereas the second penalty term resembles that of the graphical lasso for estimating the concentration matrix (Yuan and Lin, 2007; Friedman et al. 2008). When the tuning parameters, λ_{1n} and λ_{2n}, are large enough, some of the $\mu_{kj} - \mu_{lj}$ and $\Omega_{jj'}$ will estimated as zero. Further, if the following holds

for some $k \neq l$:

$$\sum_{j'=1}^{p_n} \hat{\Omega}_{jj'}(\hat{\mu}_{kj'} - \hat{\mu}_{lj'}) = 0, \tag{2.5}$$

then variable j can be considered as non-informative for differentiating classes k and l, though it may be informative for discriminating other class pairs. Moreover, if (2.5) holds for all pairs (k, l) with $1 \leq k < l \leq K$, then variable j is considered as making no contribution to the classification and can be removed from the fitted model.

One natural variation of CEDA is the doubly l_1-penalized LDA (DPL1), i.e.,

$$\max_{\omega,\mu,\Omega} l_n(\omega, \mu, \Omega) - \lambda_{1n} \sum_{j=1}^{p_n} \sum_{k=1}^{K} |\mu_{kj}| - \lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}|, \tag{2.6}$$

under the constraints $\sum_{k=1}^{K} \omega_k = 1$ and $\Omega \succ 0$. The first penalty term shrinks all class centroids towards zero, the global centroid of centered data. If all μ_{kj}'s, $k = 1, \ldots, K$, are estimated as zeros, variable j is considered as non-informative, which is in the same spirit of the "nearest shrunken centroid" method (Tibshirani et al. 2003). Note that criterion (2.6) can be considered as an improved version of the shrunken centroid method as the latter assumes the covariance matrix being diagonal while the former does not make such a strong assumption. Further, unlike (2.3), both (2.6) and the shrunken centroid method claim a variable as non-informative only when all μ_{kj}'s, $k = 1, \ldots, K$, are estimated as zeros and do not identify class-specific differentiable variables.

3 Asymptotic properties

In this section, we establish asymptotic properties of CEDA. Let $\omega = (\omega_{(1)}^\tau, \omega_K)^\tau$, where $\omega_{(1)} = (\omega_1, \ldots, \omega_{K-1})^\tau$ and $\omega_K = 1 - \sum_{k=1}^{K-1} \omega_k$. Let $\omega^* = (\omega_{(1)}^{*\tau}, \omega_K^*)^\tau$, μ^*, Ω^* and Σ^* be the true values of ω, μ, Ω and Σ, respectively. We further define two sets:

$$\begin{aligned} \mathcal{A} &= \{(j,l) : \Omega_{jl}^* \neq 0, \text{ for } j,l = 1,\ldots,p_n \text{ and } j \neq l\}, \\ \mathcal{B} &= \{(k,k',j) : \mu_{kj}^* - \mu_{k'j}^* = 0, \text{ for } 1 \leq k < k' \leq K \text{ and } j = 1,\ldots,p_n\}, \end{aligned}$$

where $\mathcal{A}$ contains the indices of off-diagonal elements in Ω^* which are truly nonzero, and $\mathcal{B}$ contains the indices of class pairs and variables that have zero mean difference.

For a symmetric matrix A, denote $\mathrm{tr}(A)$ for the trace of A, $\bar{A}$ for a diagonal matrix with the same diagonals as A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for the minimum and maximum eigenvalues of A, respectively.

Define the operator norm and the Frobenius norm, respectively, as $\|A\| = \lambda_{\max}^{1/2}(A^\tau A)$ and $\|A\|_F = \mathrm{tr}^{1/2}(A^\tau A)$. Further, we write $|\mathcal{F}|$ for the cardinality of the set $\mathcal{F}$ and $\mathcal{F}^c$ for the complement of the set $\mathcal{F}$. Let $a_n = |\mathcal{A}|$ and $b_n = K(K-1)p_n/2 - |\mathcal{B}|$. Note that a_n is the number of nonzero elements in the off-diagonal entries of Ω^*, and b_n is the number of class pair and variables that have nonzero mean differences. Finally, denote by $\tau_{ik} = I(Y_i = k)$ and $n_k = \sum_{i=1}^n \tau_{ik}$ for $i = 1,\ldots,n$ and $k = 1,\ldots,K$.

We have the following theorems that govern the asymptotic properties of CEDA. The required conditions and proofs are relegated to the Supplemental

Material.

THEOREM 1. *Let* $\hat{\omega}_{(1)}$, $\hat{\mu}$, *and* $\hat{\Omega}$ *be the maximizers defined by (2.3)-(2.4). Under conditions (A) and (B) in the Supplemental Material, if* $\log p_n/n = O(\lambda_{1n}^2)$, $\log p_n/n = O(\lambda_{2n}^2)$, *and* $(p_n+a_n)(\log p_n)^m/n = O(1)$ *for some* $m > 1$, *then we have* $\|\hat{\omega}_{(1)} - \omega^*_{(1)}\|_2^2 = O_p(n^{-1})$, $\|\hat{\mu} - \mu^*\|_2^2 = O_p(p_n \log p_n/n)$, *and* $\|\hat{\Omega} - \Omega^*\|_F^2 = O_p\{(p_n + a_n)\log p_n/n\}$.

THEOREM 2. *Under the conditions given in Theorem 1, for the maximizers of (2.3)-(2.4) satisfying* $\|\hat{\omega}_{(1)} - \omega^*_{(1)}\|_2^2 = O_p(n^{-1})$, $\|\hat{\mu} - \mu^*\|_2^2 = O_p(p_n \log p_n/n)$, $\max_{1\le j\le p_n} \|\hat{\mu}_{(j)} - \mu^*{}_{(j)}\|_2^2 = O_p(\rho_{n1})$ *for a sequence* $\rho_{n1} \to 0$, $\|\hat{\Omega} - \Omega^*\|_F^2 = O_p\{(p_n + a_n)\log p_n/n\}$, *and* $\|\hat{\Omega} - \Omega^*\|^2 = O_p(\rho_{n2})$ *for a sequence* $\rho_{n2} \to 0$, *we have the following results:*

(i) If $\log p_n/n + \rho_{n1} + \rho_{n2} = O(\lambda_{2n}^2)$, *then with probability tending to 1,* $\hat{\Omega}_{jl} = 0$ *for all* $(j,l) \in \mathcal{A}^c, j \neq l$.

(ii) If condition (C) in the Supplemental Material holds, then $\lim_{n\to\infty} P(\hat{\mathcal{B}} = \mathcal{B}) = 1$, *where* $\hat{\mathcal{B}} = \{(k,k',j) : \hat{\mu}_{kj} - \hat{\mu}_{k'j} = 0,$ *for* $1 \le k < k' \le K$ *and* $j = 1,\ldots,p_n\}$.

Theorem 1 reveals that with proper tuning parameters λ_{1n} and λ_{2n}, the CEDA estimates are consistent with certain rates of convergence. Theorem 2 shows the sparsistency of $\hat{\Omega}$ and of the fusion estimator $\hat{\mu}$, ensuring the selection consistency for the true signals among the predictors and the identification in accordance with their corresponding discriminable classes.

Further, note Theorem 1 indicates that $\hat{\mu}$ is consistent when $p_n/n = O((\log p_n)^{-m})$ with some $m > 1$, which seems restrictive. But note that there are at least p_n nonzero elements and each of them can be estimated at

best with rate $n^{-1/2}$, so the total square errors is at least of rate p_n/n. And then for high-dimensionality, we pay the price $\log p_n$. The rate decays to zero slowly, which implies that p_n can be comparable to n without violating the results in practice. The conditions here are not necessarily satisfied, but what we truly care about is the mean difference $\delta_\mu^* = \{\mu_{kj}^* - \mu_{k'j}^*, 1 \le k < k' \le K, j = 1, \ldots, p_n\}$; if δ_μ^* is sparse enough, we expect the consistency and the sparsistency hold for $p_n > n$.

To see it more clearly, we consider the binary classification problem as a special case. The following theorem establishes the asymptotic optimality of CEDA in terms of misclassification error under certain conditions on the divergence rates of b_n, p_n, a_n and $\Delta_{p_n}^2$, where $\Delta_{p_n}^2 = \delta_\mu^{*\tau} \Omega^* \delta_\mu^*$.

THEOREM 3. *In the binary case, i.e., $K = 2$, under the conditions given in Theorem 2, and assuming that*

$$c_n = \max\{\rho_{n2}^{1/2}, \frac{a_n^{1/2}}{\Delta_{p_n} n^{1/2}}, \frac{b_n^{1/2}}{\Delta_{p_n} n^{1/2}}, \frac{b_n^{1/2} \rho_{n1}^{1/2}}{\Delta_{p_n}}\} \to 0,$$

we have

(i) the conditional misclassification rate of CEDA is equal to

$$R_n = \Phi(-[1 + O_p(c_n)]\Delta_{p_n}/2),$$

where Φ is the cumulative distribution function for the standard normal distribution, and R_n is defined rigorously in the Supplemental Material;

(ii) if Δ_{p_n} is bounded, then CEDA is asymptotically optimal and

$$\frac{R_n}{R_{\text{OPT}}} - 1 = O_p(c_n),$$

where $R_{\text{OPT}} = \Phi(-\Delta_{p_n}/2)$ denotes the misclassification rate of the optimal classification rule (Anderson 2003);

(iii) if $\Delta_{p_n} \to \infty$, then CEDA is asymptotically sub-optimal, i.e.,

$$R_n - R_{\text{OPT}} \xrightarrow{P} 0;$$

(iv) if $\Delta_{p_n} \to \infty$ and $c_n \Delta_{p_n}^2 \to 0$, then CEDA is asymptotically optimal.

4 Implementation and tuning parameter selection

We propose an algorithm to implement the proposed method and also propose a procedure to select the tuning parameters λ_{1n} and λ_{2n}.

Firstly, we note that $\hat{\omega}_k = \sum_{i=1}^n I(y_i = k)/n$, for $k = 1, \ldots, K$, whereas the estimators of μ and Ω can be obtained through an iterative algorithm: we fix μ and estimate Ω, then we fix the estimated Ω and estimate μ; we iterate between these two steps until the algorithm converges. Since the value of the objective function (2.3) decreases over iterations, convergence is guaranteed.

When μ is fixed, to maximize Q_n with respect to Ω, it suffices to maximize

$$Q_1(\Omega) = \log|\Omega| - \text{tr}(S(\mu)\Omega) - \frac{1}{2}\lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}| \tag{4.1}$$

over all non-negative definite matrices Ω for a known covariance matrix $S(\mu)$. Note that it is similar to the problem of estimating sparse graphs. Hence, we can apply the graphical lasso algorithm proposed by Friedman et al. (2008) to efficiently solve for Ω.

When Ω is fixed, to maximize Q_n with respect to μ, it suffices to minimize

$$n^{-1} \sum_{i=1}^n \sum_{k=1}^K I(y_i = k)(x_i - \mu_k)^\tau \Omega (x_i - \mu_k) + \frac{1}{2}\lambda_{1n} \sum_{j=1}^{p_n} \sum_{1 \leq k < k' \leq K} |\mu_{kj} - \mu_{k'j}|. \tag{4.2}$$

Note it is challenging to directly minimize (4.2) with respect to μ due to the fusion penalty. We apply the local quadratic approximation (Fan and Li 2001) to convert the minimization in (4.2) into a generalized ridge problem. Specifically, we approximate

$$|\mu_{kj}^{(t+1)} - \mu_{k'j}^{(t+1)}| \approx \frac{(\mu_{kj}^{(t+1)} - \mu_{k'j}^{(t+1)})^2}{2|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|} + \frac{1}{2}|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|,$$

where t is the iteration index used to denote iterations of the local quadratic approximation. Consequently, we only need to consider the following objective function

$$\begin{aligned} Q_2(\mu) &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i = k)(x_i - \mu_k)^\tau \Omega (x_i - \mu_k) \\ &\quad + \frac{1}{2}\lambda_{1n}\sum_{j=1}^{p_n}\sum_{1\le k<k'\le K} \frac{(\mu_{kj} - \mu_{k'j})^2}{2|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|}, \end{aligned} \tag{4.3}$$

and thus, $\mu^{(t+1)} = \arg\min_\mu Q_2(\mu)$.

Overall, the algorithm proceeds as follows:

1. Initialization: Initialize $\mu^{(0)}$ with some plausible values, and set $s = 1$.

2. Update of Ω: For iteration s, apply the graphical lasso algorithm to maximize (4.1) with μ replaced by $\mu^{(s-1)}$ and obtain $\Omega^{(s)}$.

3. Update of μ: With Ω replaced by $\Omega^{(s)}$, iteratively minimize the generalized ridge criterion (4.3) until $\sum_{j=1}^{p_n}\sum_{k=1}^{K}|\mu_{kj}^{(t+1)} - \mu_{kj}^{(t)}| / \sum_{j=1}^{p_n}\sum_{k=1}^{K}|\mu_{kj}^{(t)}|$ is small enough to obtain $\mu^{(s)}$.

4. Stopping criterion: If $|Q_n(\hat{\omega}, \mu^{(s)}, \Omega^{(s)}) - Q_n(\hat{\omega}, \mu^{(s-1)}, \Omega^{(s-1)})|$ is small enough, stop the algorithm. Otherwise, set $s \leftarrow s + 1$ and go back to Step 2.

In terms of selecting the tuning parameters λ_{1n} and λ_{2n}, we follow the suggestion in Wang et al. (2007) and use a BIC-type criterion:

$$BIC(\lambda_{1n}, \lambda_{2n}) = -2nl_n(\hat{\omega}, \hat{\mu}, \hat{\Omega}) + (K - 1 + d_{\hat{\mu}} + d_{\hat{\Omega}}) \log(n), \tag{4.4}$$

where $d_{\hat{\mu}}$ is the number of distinct nonzero elements in $\hat{\mu}$ and $d_{\hat{\Omega}}$ is the number of nonzero elements in $\hat{\Omega}$.

5 Simulation studies

In this section, we use simulation studies to assess the finite sample performance of the proposed CEDA method, and compare it with two major competing methods - the LDAPF method proposed by Guo (2010) and the DPL1 method defined in (2.6). The LDAPF method is a special case of our CEDA method assuming the covariance matrix being diagonal.

Example 1 Consider a three-class scenario with a total of $p = 210$ variables, generated according to the following mechanism: the first 10 variables are independently distributed $N(\mu_{kj}, 1)$ for class k, whereas the remaining 200 variables are i.i.d. from $N(0, 1)$ for all three classes. Table 1 gives the means for the first 10 variables. For example, in class 1, variables 1-5 all have the same mean value 0, and variables 6-10 all have the same mean value 1.5.

EXAMPLE 2 The true model is the same as that in Example 1 except that the covariance matrix is non-diagonal. Specifically, we consider the AR(1) correlation structure with auto-correlation coefficient 0.6 for variables 1-5 and variables 6-10, respectively. Variables 1-5 are independent of variables 6-10, and both groups are independent of the remaining 200 variables.

EXAMPLE 3 The true model is the same as that in Example 1 except that variable 5 has different means from variables 1-4 and the correlation structure among variables 1-10 is also different from those in Examples 1 and 2. Specifically, the means of variable 5 are respectively -0.5, 2, and -2.5 in the 3 classes. Variables 1-5 now have an exchangeable correlation structure with coefficient 0.5. Variables 6-10 are correlated with the same structure but independent of variables 1-5. Table 1 gives the means for the first 10 variables.

Table 1: Means of the informative variables in simulated examples 1-3

Example	Variables	Class 1	Class 2	Class 3
1 & 2	1-5	0	0	-2.5
	6-10	1.5	-1.5	-1.5
3	1-4	0	0	-2.5
	5	-0.5	2	-2.5
	6-10	1.5	-1.5	-1.5

Using criterion (2.1), we note that only the first 10 variables are informative in each simulation example. Moreover, in Examples 1 and 2, a variable is informative for separating a pair of classes if it has unequal means for the

corresponding classes. For example, variables 1-5 are informative for separating classes 1 and 3 or classes 2 and 3, but non-informative for separating classes 1 and 2, similarly for variables 6-10. While for Example 3, it is a little tricky to identify the informative variables for discriminating classes 1 and 2. For example, variable 1 has equal mean effects for classes 1 and 2, but in fact it does contribute to the classification through its correlation with the informative variable 5, just as what Figure 1 has illustrated. Therefore, unlike in Examples 1 and 2, variables 1-5 are all informative for separating classes 1 and 2.

In each example, we generate 50 data sets, each consisting of $n_1 = n_2 = n_3 = 50$ training and test samples. We then apply each method to the training data and record the average misclassification error rate (ER) evaluated on the testing data, the average proportion of incorrectly removed informative variables, i.e., the false negative rate (FN), the average proportion of incorrectly selected non-informative variables, i.e., the false positive rate (FP), and the average model size (MS).

Table 5 summarizes the misclassification error rates and the variable selection results of the three methods over 50 replications. We can see that in all examples, CEDA significantly outperforms LDAPF and DPL1 in terms of classification accuracy. In terms of variable selection, all three methods are effective at identifying the informative variables, but CEDA is more effective than LDAPF and DPL1 in removing non-informative variables. It is also worth to note that CEDA achieves the competitive prediction accuracy with

a much smaller model, comparing with the other two methods; the standard deviation of the model size is also much smaller for the CEDA method.

If a variable is non-informative for discriminating a pair of classes, and the corresponding estimated parameters satisfy equation (2.5), we consider it as correct "fusion". Table 5 summarizes the fusion results for all the examples. Specifically, each row in the table presents the average proportion of fused variables out of the five for separating the corresponding pair of classes. For example, the first row indicates that for CEDA, on average 99.2% of the fist five variables are fused for classes 1 and 2. Note that 100% is the optimal value except for the fifth row of the table as variables 1-5 are in fact informative for separating classes 1 and 2 in Example 3, and thus 0% should be the optimal value for the fifth row. From Table 5 we can see that CEDA dominates the LDAPF method in terms of correctly fusing variables for separating a specific pair of classes. We also note that DPL1 never fuses any of the first 10 variables judged by the criterion (2.5). This phenomenon is understandable as DPL1 only penalizes the individual μ_{kj}'s, not the pairwise differences; thus a variable can only be fused if all μ_{kj}, $k = 1, \ldots, K$ are estimated as zero, but clearly it is not a favorable estimate for the first 10 variables as the true class means are different.

Table 2: Misclassification error rates and variable selection results for Examples 1-3. Each table cell represents the average(SD) over 50 repetitions. "ER" is the average misclassification error rate on the test data set, "FN" is the average false negative rate, "FP" is the average false-positive rate, and "MS" is the average model size.

Example	Method	ER(%)	FN(%)	FP(%)	MS
1	CEDA	0.16(0.35)	0.00(0.00)	0.29(0.50)	10.58(0.99)
	LDAPF	0.27(0.43)	0.00(0.00)	7.74(6.43)	25.48(12.86)
	DPL1	9.65(3.35)	0.00(0.00)	59.47(8.45)	128.94(16.90)
2	CEDA	3.96(1.32)	0.00(0.00)	1.20(0.80)	12.40(1.60)
	LDAPF	4.09(1.26)	0.00(0.00)	10.39(10.67)	30.78(21.35)
	DPL1	14.09(3.08)	0.00(0.00)	89.95(9.49)	189.90(18.99)
3	CEDA	1.84(0.98)	0.00(0.00)	0.48(0.49)	10.96(0.99)
	LDAPF	8.01(2.05)	0.00(0.00)	9.23(6.45)	28.46(12.90)
	DPL1	2.47(1.14)	0.00(0.00)	64.83(10.38)	139.66(20.77)

6 Application to the kidney transplant rejection and tissue injury data

The kidney transplant rejection and tissue injury data set in Flechner et al. (2004) consists of 62 tissue samples from kidney transplant patients, including normal donor kidneys (C) (17 samples), well-functioning transplants without rejection (TX) (19 samples), kidneys undergoing acute rejection (AR) (13 samples), and transplants with renal dysfunction without rejection (NR) (13 samples). Each sample is described by 12,625 genes from kidney biopsies and peripheral blood lymphocytes. Distinguishing these four types of patients is

Table 3: Pairwise class fusion results for Examples 1-3. "Pair" corresponds to a pair of indiscriminable classes pairs for the variables in the corresponding row (except for the fifth row). For example, the fist row indicates that variables 1-5 are non-informative for separating classes 1 and 2. The numbers in the following columns give the proportions of variables in the set that are identified as non-informative for separating a given pair of classes by each method. The optimal value is 100% in each case except for the fifth row, where the optimal value should be 0%. All results are averaged over 50 repetitions with the corresponding standard deviations in the parentheses.

Example	Variables	Pair	CEDA(%)	LDAPF(%)	DPL1(%)
1	1-5	1/2	99.20(3.96)	90.00(14.14)	0.00(0.00)
	6-10	2/3	99.60(2.83)	85.60(16.68)	0.00(0.00)
2	1-5	1/2	96.80(13.01)	82.80(20.21)	0.00(0.00)
	6-10	2/3	98.00(7.28)	85.60(20.22)	0.00(0.00)
3	1-5	1/2	1.20(3.28)	33.00(9.95)	0.00(0.00)
	6-10	2/3	99.60(2.83)	89.20(17.71)	0.00(0.00)

crucial to balance the need for immunosuppression to prevent rejection, while minimizing drug-induced toxicities.

Before applying our method, we conduct a prescreening step as commonly done in literature. Specifically, we pre-select a subset of genes following Guo et al. (2010) according to their variances. In practice, genes with largest and smallest variabilities are generally considered to be potentially most relevant to biological functions. We select 100 genes with largest variances and 100 genes with smallest variances from the 12,625 genes. The selection does not use any class label information. Then, we center the obtained 200 genes before classification.

To assess the performance, we randomly split the data set into training and test sets with ratio 2:1. We estimate and select the genes on the training data set and then, evaluate the classification accuracy on the test data set. This procedure is repeated 100 times. Figure 2 summarizes the classification accuracy using boxplots for the CEDA, the LDAPF and the DPL1 methods, and suggests that CEDA performs the best and DPL1 performs the worst.

To assess variable selection, we count the selected times of each gene based on 100 random splits. Then we choose the 25 most "informative" genes according to their frequency. There are 19 most "informative" genes selected by all three methods and besides these 19 common genes, the CEDA method selected the following genes as the most "informative" genes: HCFC1, PLIN2, LOC646347, IDS, SPAG5, and TIGR(HG4518-HT4921), some of which are significantly relevant to renal functions. For example, the HCFC1 gene, as a member of the host cell factor family, was reported in Wilson et al. (1995) to be highly expressed in fetal tissues and the adult kidney; the expression of PLIN2 has been shown as a predictor of cancer-specific survival in clear cell renal carcinoma (Yao et al. 2007); SPAG5 is highly expressed in human normal kidneys (Chang et al. 2001) while the level of expression is extremely lower in hgn/hgn kidneys than in normal kidneys (Suzuki et al. 2006).

Further, CEDA reveals that not all the selected 19 most "informative" genes are informative for discriminating every pair of classes. For example, Figure 3 shows gene AGGF1, reported to have strong protein expression in blood vessels embedded in kidney tissues (Fan et al. 2009), cannot discrim-

inate class AR from class NR but is informative for other class pairs; gene GRINA, which plays a major role in gentamicin ototoxicity (Leung et al. 2004) and in 1,25$(OH)_2$ D_3 synthesis (Parisi et al. 2010), cannot separate classes C, AR and NR; gene RFNG which is strongly expressed in the kidney (Challen et al. 2006) cannot discriminate class C from class AR. Furthermore, though some of genes have the same means across different classes, they are informative in classification via correlations with other informative genes. For example, gene AGGF1 can discriminate class C from classes AR and NR, though it has the same means within these three classes based on Figure 4; gene COMT is informative to separate any pair of classes, especially the classes AR and NR; gene RFNG has contribution to discriminate class NR from classes C and AR via the correlation.

In summary, our proposed method helps gain additional insight. Firstly, it identifies new genes that are relevant to renal functions, which is of biological significance. Secondly, by utilizing the underlying covariance structures between genes, the method elucidates the impact of genes on differentiating particular renal functional classes, a crucial step in the development of gene therapy.

7 Discussion

There exists an intrinsic relationship between LDA and multinomial logistic regression, as $\hat{\Omega}(\hat{\mu}_k - \hat{\mu}_l)$, obtained from (2.3), can be viewed as the differences of estimated regression coefficients in multinomial regression models. How-

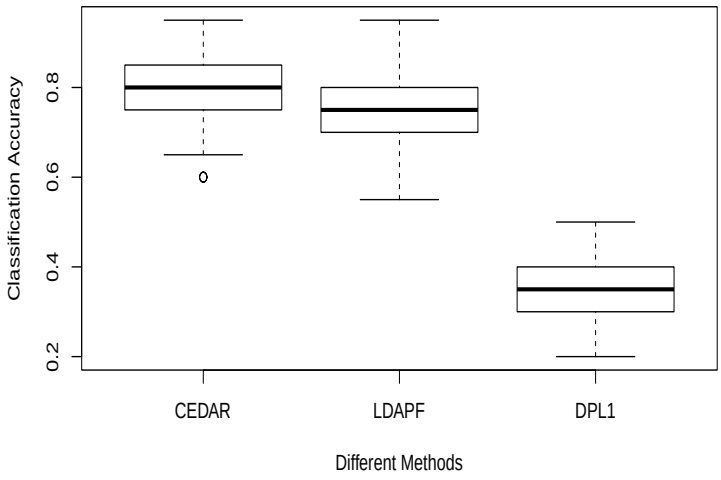

Figure 2: Classification accuracies of the three methods on the kidney transplant rejection and tissue injury data set

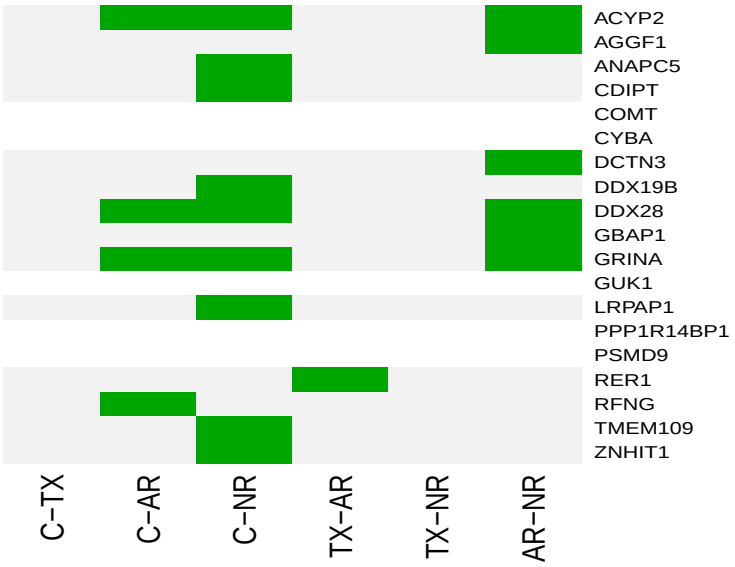

Figure 3: Pairwise class fusion results for the CEDA method with the 19 most "informative" genes selected in the kidney transplant rejection and tissue injury data set. Each row corresponds to a gene. Each column corresponds to a class pair. A green/dark spot indicates that the corresponding gene is non-informative for separating the corresponding pair of classes.

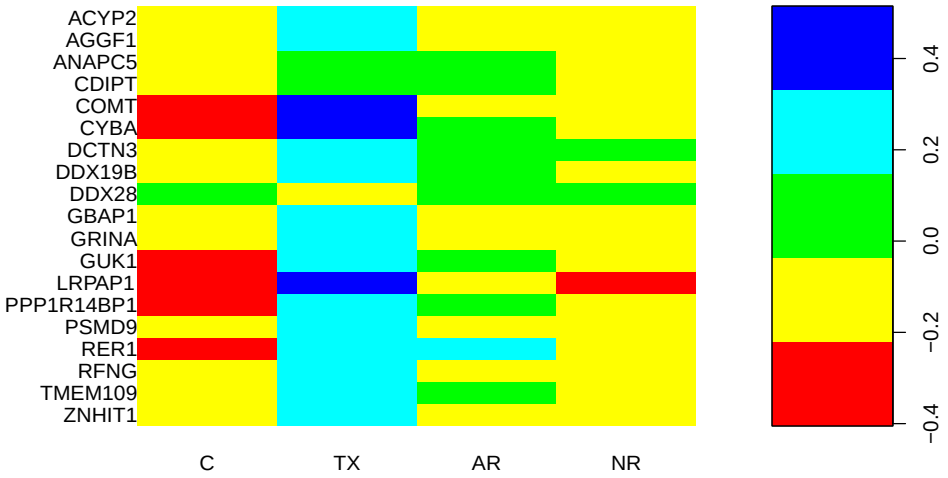

Figure 4: The left is the heatmap of the estimated centroids for the 19 most "informative" genes selected in the kidney transplant rejection and tissue injury data set. Rows correspond to genes and columns to classes. The right is the color key.

ever, as opposed to a multinomial logistic regression that cannot effectively utilize the covariance between covariates, we have proposed a covariance-enhanced classification method for simultaneously selecting informative variables and identifying the corresponding discriminable classes. In particular, our method penalizes the off-diagonal elements of the concentration matrix and the difference between class means for each pair of classes and for each variable, which allows one to identify and remove non-informative variables for selected pair of classes when both mean effect and covariance effect are considered. This helps to improve the interpretation of the effect of a particular variable on differentiating different classes. Further, our method enjoys plausible theoretical and numerical properties and performs well in real data analysis.

Possible extensions include the discrimination in the case with different covariance matrices for different classes, and that with non-Gaussian data. Other opportunities also include the applications of the proposed method to problems such as clustering, which is ongoing.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis.* Third edition. Wiley-Interscience.

[2] Bickel, P.J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli*, **10**, 989-1010.

[3] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices, *Annals of Statistics*, **36**, 199-227.

[4] Cai, T. and Liu, W.D. (2011). A direct estimation approach to sparse linear discriminant analysis, *Journal of the American Statistical Association*, **496**, 1566-1577.

[5] Chang, M.S., Huang, C.J., Chen, M.L., Chen, S.T., Fan, C.C., Chu, J.M., Lin, W.C., and Yang, Y.C. (2001). Cloning and characterization of hMAP126, a new member of mitotic spindle-associated proteins, *Biochemical and Biophysical Research Communications*, **287**, 116-121.

[6] Challen, G.A., Bertoncello, I., Deane, J.A., Ricardo, S.D., and Little, M.H. (2006). Kidney Side Population Reveals Multilineage Potential and Renal Functional Capacity but also Cellular Heterogeneity, *Journal of the American Society of Nephrology*, **17**, 1896-1912.

[7] Clemmensen, L., Hastie, T., Witten, D., and Ersboll, B. (2011). Sparse discriminant analysis, *Technometrics*, **53**, 406-413.

[8] Fan, C., Quyang, P., Timur, A.A., He, P., You, S.A., Hu, Y., Ke, T., Driscoll, D.J., Chen, Q., and Wang, Q.K. (2009). Novel Roles of GATA1 in Regulation of Angiogenic Factor AGGF1 and Endothelial Cell Function, *The Journal of Biological Chemistry*, **284**, 23331-23343.

[9] Fan, J.Q. and Fan, Y.Y. (2008). High-dimensional classification using features annealed independence rules, *Annals of Statistics*, **36**, 2605-2637.

[10] Fan, J.Q. and Li, R.Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

[11] Flechner, S.M., Kurian, S.M., Head, S.R., Sharp, S.M., Whisenant, T.C., Zhang, J., Chismar, J.D., Horvath, S., Mondala, T., Gilmartin, T., Cook, D.J., Kay, S.A., Walker, J.R., and Salomon, D.R. (2004). Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes, *American Journal of Transplantation*, **4**, 1475-1489.

[12] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9**, 432-441.

[13] Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis, *Biostatistics*, **11**, 599-608.

[14] Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering, *Biometrics*, **66**, 793-804.

[15] Guo, Y.Q., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, **8**, 86-100.

[16] Ledoux, M. and Talagrand, M. (1991). *Prbability in Banach Spaces: Isoperimetry and Processes.* Springer, Berlin.

[17] Leung, J.C., Marphis, T., Craver, R.D., and Silverstein, D.M. (2004). Altered NMDA receptor expression in renal toxicity: Protection with a receptor antagonist. *Kidney International*, **66**, 167-176.

[18] Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions, *Biometrika*, **99**, 29-42.

[19] Parisi, E., Bozic, M., Ibarz, M., Panizo, S., Valcheva, P., Coll, B., Fernández, E., and Valdivielso, J.M. (2010). Sustained activation of renal N-methyl-D-aspartate receptors decreases vitamin D synthesis: a possible role for glutamate on the onset of secondary HPT, *American Journal of Physiology - Endocrinology and Metabolism*, **299**, E825-E831.

[20] Qiao, Z., Zhou, L., and Huang, J.Z. (2008). Sparse linear discriminant analysis with applications to high dimensional low sample size data, *IAENG International Journal of Applied Mathematics*, **39**, 48-60.

[21] Rinaldo, A. (2009). Properties and refinements of the fused lasso, *Annals of Statistics*, **37**, 2922-2952.

[22] Shao, J., Wang, Y.Z., Deng, X.W., and Wang, S.J. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data, *Annals of Statistics*, **39**, 1241-1265.

[23] Suzuki, H., Yagi, M., and Suzuki, K. (2006) Duplicated insertion mutation in the microtubule-associated protein Spag5 (astrin/MAP126) and defective proliferation of immature Sertoli cells in rat hypogonadic (hgn/hgn) testes, *Reproduction*, **132**, 79-93.

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B*, **58**, 267-288.

[25] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567-6572.

[26] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104-117.

[27] Wang, H., Li, R. and Tsai, C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553-568.

[28] Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier, *Bioinformatics*, **23**, 972-979.

[29] Witten, D.M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant, *Journal of the Royal Statistical Society: Series B*, **73**, 753-772.

[30] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model, *Biometrika*, **94**, 19-35.

8 Supplemental Material

In this section the proofs of theorems are given. We first introduce the following regularity conditions:

(A) There are positive constants κ_1 and κ_2 such that $\kappa_1 < \lambda_{\min}(\Sigma^*) \leq \lambda_{\max}(\Sigma^*) < \kappa_2$ for all n.

(B) $\min_{1\leq k\leq K} n_k/n = O_p(1)$.

(C) For some $\eta > 0$,

(i). $\frac{\lambda_{1n}p_n^{1/2}}{b_{\max}^{*1/2}} \to \infty$, $\frac{\lambda_{1n}p_n^{1/2}}{\left\{b_{\max}^* \log\left(\frac{K(K-1)p_n}{2}-b_n\right)\right\}^{1/2}} > 1+\eta$ and $\alpha_n^{\max} = o_p(\lambda_{1n}p_n^{1/2})$,

(ii). $\frac{\alpha_n^{\min}}{b_{\max}^{*1/2}} \to \infty$, $\frac{\alpha_n^{\min}}{(b_{\max}^* \log b_n)^{1/2}} > 1+\eta$ and $\lambda_{1n} < \frac{\alpha_n^{\min}}{4\kappa_2 p_n^{1/2}(K-1)}$,

where $b_{\max}^* = \max_{1\leq j\leq p_n} \sigma_{jj}^*$, $\alpha_n^{\max} = \max_{\mathcal{B}} \left|\sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right) \sum_{l=1}^K \tau_{il}\mu_{lj}^*\right|$, and $\alpha_n^{\min} = \min_{\mathcal{B}^c} \left|\sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right) \sum_{l=1}^K \tau_{il}\mu_{lj}^*\right|$.

Condition (A) bounds the eigenvalues of the covariance matrix Σ^* uniformly, and condition (B) implies that the K samples are of comparable sizes. Both are the commonly used conditions in the high dimensional setting (see Cai and Liu 2011), which facilitates the proof for consistency. Condition (C) is analogous to the conditions in Theorem 2.3 of Rinaldo (2009), which is used for the proof of sparsistency. From the proof, we can see that $\alpha_n^{\max}$ and $\alpha_n^{\min}$ are related to the magnitude of mean difference $\delta_\mu^* = \{\mu_{kj}^* - \mu_{k'j}^*, 1 \leq k < k' \leq K, j = 1, \ldots, p_n\}$, whose asymptotic behavior

determines whether recovery of the true mean effect obtained. In particular, if $\alpha_n^{\min}$ vanishes at a rate faster than $1/b_{\max}^{*1/2}$, then no recovery is possible.

Proof of Theorem 1. The proof is summarized in the following three steps. First, we prove $Q_n(\omega^*, \mu^*, \Omega^*) \geq Q_n(\omega, \mu^*, \Omega^*)$ for $\|\hat{\omega}_{(1)} - \omega_{(1)}^*\|_2^2 = O_p(n^{-1})$. In Step 2, we show that $Q_n(\omega, \mu^*, \Omega^*) \geq Q_n(\omega, \mu^*, \Omega)$ for $\|\hat{\Omega} - \Omega^*\|_F^2 = O_p\{(p_n + a_n)\log p_n/n\}$. In Step 3, we prove that $Q_n(\omega, \mu^*, \Omega) \geq Q_n(\omega, \mu, \Omega)$ for $\|\hat{\mu} - \mu^*\|_2^2 = O_p(p_n \log p_n/n)$. The following are the details.

Step 1. Let $\Delta_{\omega_{(1)}} = \omega_{(1)} - \omega_{(1)}^*$, and $h(\omega_{(1)}) = \sum_{i=1}^n \sum_{k=1}^K \tau_{ik} \log \omega_k$, where $\omega_K = 1 - \sum_{k=1}^{K-1} \omega_k$. We denote by $J_\omega = (\delta_1, \ldots, \delta_K)^\tau$ the Jacobian matrix, where $\delta_k (1 \leq k < K)$ is a $(K-1)$-dimensional unit vector with the kth component being 1, and δ_K is a $(K-1)$-dimensional vector of ones. An application of Taylor expansion yields

$$\begin{aligned}
& Q_n(\omega, \mu^*, \Omega^*) - Q_n(\omega^*, \mu^*, \Omega^*) \\
= \; & \frac{1}{n} J_\omega^\tau \frac{\partial h(\omega_{(1)}^*)}{\partial \omega} \Delta_{\omega_{(1)}} - \frac{1}{2} \Delta_{\omega_{(1)}}^\tau J_\omega^\tau \left\{ -\frac{1}{n} \frac{\partial^2 h(\omega_{(1)}^*)}{\partial \omega \partial \omega^\tau} \right\} J_\omega \Delta_{\omega_{(1)}} \\
& + o_p \left(\Delta_{\omega_{(1)}}^\tau J_\omega^\tau \left\{ -\frac{1}{n} \frac{\partial^2 h(\omega_{(1)}^*)}{\partial \omega \partial \omega^\tau} \right\} J_\omega \Delta_{\omega_{(1)}} \right) \\
\triangleq \; & A_1 - A_2 + A_3.
\end{aligned}$$

Note that $n^{-1} \sum_{i=1}^n \{\tau_{ik} \omega_k^{*-1} - \tau_{iK} \omega_K^{*-1}\} = o_p(1)$ because $E\tau_{ik} = \omega_k^*$ for $k = 1, \ldots, K$. Consequently, we have

$$\begin{aligned}
A_1 & \leq n^{-1/2} O_p(1) \|\Delta_{\omega_{(1)}}\|_1 \\
& \leq (K-1)^{1/2} O_p(n^{-1/2}) \|\Delta_{\omega_{(1)}}\|_2.
\end{aligned}$$

Further, since $n^{-1}\sum_{i=1}^{n}\tau_{ik}\omega_k^{*-2} \xrightarrow{P} \omega_k^{*-1}$ for $k = 1, \ldots, K$, we have

$$J_\omega^\tau \left\{ -\frac{1}{n}\frac{\partial^2 h(\omega_{(1)}^*)}{\partial\omega\partial\omega^\tau} \right\} J_\omega \xrightarrow{P} J_\omega^\tau H J_\omega > 0,$$

where H is a $K \times K$ diagonal matrix with the kth element ω_k^{*-1}. Hence,

$$A_2 \geq \frac{1}{2} O_p(1) \|\Delta_{\omega_{(1)}}\|_2^2,$$

implying that A_2 dominates both A_1 and A_3 uniformly in $\|\hat{\omega}_{(1)} - \omega_{(1)}^*\|_2^2 = O_p(n^{-1})$. Therefore, $Q_n(\omega^*, \mu^*, \Omega^*) \geq Q_n(\omega, \mu^*, \Omega^*)$ for $\|\hat{\omega}_{(1)} - \omega_{(1)}^*\|_2^2 = O_p(n^{-1})$.

Step 2. Let $\Delta_\Omega = \Omega - \Omega$ and $S = S(\mu^*)$. Consider the difference

$$Q_n(\omega, \mu^*, \Omega) - Q_n(\omega, \mu^*, \Omega^*) = B_1 - B_2 - B_3,$$

where

$$\begin{aligned} B_1 &= 2^{-1}\left(\log|\Omega| - \log|\Omega^*|\right) - 2^{-1}\mathrm{tr}(S\Delta_\Omega), \\ B_2 &= \lambda_{2n} \sum_{(j,l)\in\mathcal{A}^c, j\neq l} (|\Omega_{jl}| - |\Omega_{jl}^*|), \\ B_3 &= \lambda_{2n} \sum_{(j,l)\in\mathcal{A}} (|\Omega_{jl}| - |\Omega_{jl}^*|). \end{aligned}$$

An application of Taylor expansion with the integral remainder yields that

$$\log|\Omega| - \log|\Omega^*| = \mathrm{tr}(\Sigma^*\Delta_\Omega) - \vec{\Delta}_\Omega^\tau \left\{ \int_0^1 (1-v)\Omega_v^{-1} \otimes \Omega_v^{-1} dv \right\} \vec{\Delta}_\Omega,$$

where $\Omega_v = \Omega^* + v\Delta_\Omega$ with $0 \leq v \leq 1$, $\vec{\Delta}_\Omega$ is the vectorization of Δ_Ω, and $\otimes$ is the Kronecker product. Therefore, B_1 can be written as $B_1 = -2^{-1}(I_1 + I_2)$,

where

$$
\begin{aligned}
I_1 &= \mathrm{tr}\left((S-\Sigma^*)\Delta_\Omega\right), \\
I_2 &= \vec{\Delta}_\Omega^\tau \left\{ \int_0^1 (1-v)\Omega_v^{-1} \otimes \Omega_v^{-1} dv \right\} \vec{\Delta}_\Omega.
\end{aligned}
$$

First consider I_1. Let s_{jl}, σ^*_{jl}, and $\Delta_{\Omega jl}$ be respectively the (j,l)th element of S, Σ^* and Δ_Ω. Denote by $\mathcal{C} = \{(j,j) : j = 1, \ldots, p_n\}$. Then, it is clear that $|I_1| \le I_{11} + I_{12}$, where

$$
\begin{aligned}
I_{11} &= |\sum_{(j,l)\in\mathcal{A}\cup\mathcal{C}} (s_{jl}-\sigma^*_{jl})\Delta_{\Omega jl}|, \\
I_{12} &= |\sum_{(j,l)\in\mathcal{A}^c, j\neq l} (s_{jl}-\sigma^*_{jl})\Delta_{\Omega jl}|.
\end{aligned}
$$

Let $z_i = \sum_{k=1}^K \tau_{ik}(x_i - \mu^*_k)$ for $i = 1, \ldots, n$. By the assumption, $z_i = (z_{i1}, \ldots, z_{ip})^\tau$'s are i.i.d. p-variate normal random variables with mean 0 and covariance matrix Σ^*. Note that $s_{jl} = n^{-1}\sum_{i=1}^n z_{ij}z_{il}$. Using Lemma 3 in Bickel and Levina (2008), we have

$$
\begin{aligned}
I_{11} &\le (p_n + a_n)^{1/2} \max_{(j,l)\in\mathcal{A}\cup\mathcal{C}} |s_{jl}-\sigma^*_{jl}| \cdot \|\Delta_\Omega\|_F \\
&\le O_p(\{(p_n+a_n)\log p_n/n\}^{1/2}) \cdot \|\Delta_\Omega\|_F \\
&= O_p((p_n+a_n)\log p_n/n).
\end{aligned}
$$

Consider $B_2 - I_{12}$ for penalties. Note that $\Delta_{\Omega jl} = \Omega_{jl}$ for all $(j,l) \in \mathcal{A}^c$, $j \neq l$. Invoking Lemma 3 in Bickel and Levina (2008) again, we have

$$
\begin{aligned}
B_2 - I_{12} &\ge \lambda_{2n} \sum_{(j,l)\in\mathcal{A}^c, j\neq l} |\Omega_{jl}| - \max_{(j,l)} |s_{jl}-\sigma^*_{jl}| \sum_{(j,l)\in\mathcal{A}^c, j\neq l} |\Delta_{\Omega jl}| \\
&\ge \sum_{(j,l)\in\mathcal{A}^c, j\neq l} [\lambda_{2n} - O_p(\{\log p_n/n\}^{1/2})]|\Omega_{jl}| \\
&\ge 0
\end{aligned}
$$

for $\lambda_{2n}^2 = O(\log p_n/n)$. For the term B_3, we have

$$\begin{aligned} B_3 &= \lambda_{2n} \sum_{(j,l)\in\mathcal{A}} (|\Omega_{jl}| - |\Omega_{jl}^*|) \\ &\leq \lambda_{2n} \sum_{(j,l)\in\mathcal{A}} |\Delta_{\Omega jl}| \\ &\leq \lambda_{2n} a_n^{1/2} \|\Delta_\Omega\|_F \\ &= O_p((p_n + a_n)\log p_n/n). \end{aligned}$$

Finally, we bound I_2. Recall that $\lambda_{\min}(M) = \min_{\|x\|=1} x^\tau M x$ for any symmetric matrix M. Then, under condition (A), we have

$$\begin{aligned} I_2 &\geq \int_0^1 (1-v) \min_{0\leq v\leq 1} \lambda_{\min}(\Omega_v^{-1} \otimes \Omega_v^{-1}) dv \cdot \|\vec{\Delta}_\Omega\|_2^2 \\ &= \|\vec{\Delta}_\Omega\|_2^2/2 \cdot \min_{0\leq v\leq 1} \lambda_{\max}^{-2}(\Omega_v) \\ &\geq \|\vec{\Delta}_\Omega\|_2^2/2 \cdot (\kappa_1 + o(1))^{-2} \\ &= C_1(p_n + a_n)\log p_n/n, \end{aligned}$$

for a large constant C_1. To derive the above inequality, we have used $\|\Delta_\Omega\| \leq \|\Delta_\Omega\|_F = O((\log p_n)^{(1-m)/2}) = o(1)$ by our assumption. Therefore, I_2 dominates both I_{11} and B_3 with a large constant C_1. With $B_2 - I_{12} \geq 0$, this completes the proof of the Step 2.

Step 3. Let $\Delta_{\mu_k} = (\Delta_{\mu_{k1}}, \ldots, \Delta_{\mu_{kp_n}})^\tau = \mu_k - \mu_k^*$, for $k = 1, \ldots, K$, and $\Delta_\mu = \mu - \mu^*$. Then, for each $1 \leq k \leq K$, $\Delta_{\mu_k} = (I_{p_n} \otimes e_k^\tau)\Delta_\mu$, where I_{p_n} is a $p_n \times p_n$ identity matrix and e_k is a K-dimensional unit vector with kth component 1. For the sake of simplicity, let $z_i = \sum_{k=1}^K \tau_{ik}(x_i - \mu_k^*)$ and $E_i = \sum_{k=1}^K \tau_{ik}(I_{p_n} \otimes e_k^\tau)$, for $i = 1, \ldots, n$. Consider the difference

$$Q_n(\omega, \mu, \Omega) - Q_n(\omega, \mu^*, \Omega) = I_1' - I_2' + I_3'$$

where

$$\begin{aligned} I_1' &= n^{-1}\sum_{i=1}^{n} z_i^\tau \Omega E_i \Delta_\mu, \\ I_2' &= (2n)^{-1}\sum_{i=1}^{n} \Delta_\mu^\tau E_i^\tau \Omega E_i \Delta_\mu^\tau, \\ I_3' &= -\lambda_{1n}\sum_{j=1}^{p_n}\sum_{1\le k<k'\le K}\left[|\mu_{kj}-\mu_{k'j}| - |\mu_{kj}^*-\mu_{k'j}^*|\right]. \end{aligned}$$

Let $\Delta_\mu^{(s)}$ be the sth component of Δ_μ, and δ_s' be a (Kp_n)-dimensional unit vector with sth component 1, for $s = 1, \ldots, Kp_n$. Then, it can be seen that $|I_1'| = \sum_{s=1}^{Kp_n} \eta_s \Delta_\mu^{(s)}$, where

$$\eta_s = n^{-1}\sum_{i=1}^{n} z_i^\tau \Omega E_i \delta_s',$$

for $s = 1, \ldots, Kp_n$. Now, consider the event $\mathcal{F} = \bigcap_{s=1}^{Kp_n}\{|\eta_s| \le \lambda_{1n}\}$. Since $\|\Omega - \Omega^*\| = o_p(1)$, we have $\|\Omega\Sigma^* - I_{p_n}\| = o_p(1)$ by condition (A). Thus, $\|\Omega\Sigma^*\Omega - \Omega^*\| = \|(\Omega\Sigma - I_{p_n})(\Omega - \Omega^*)\| = o_p(1)$. Consequently,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} \delta_s'^\tau E_i^\tau \Omega\Sigma^*\Omega E_i \delta_s' &= n^{-1}\sum_{i=1}^{n} \delta_s'^\tau E_i^\tau \Omega^* E_i \delta_s' + o_p(1) \\ &\triangleq M_s + o_p(1). \end{aligned}$$

Therefore, using the probability bound on the tail of the standard Gaussian distribution, we know that

$$\begin{aligned} P(\mathcal{F}^c) &\le \sum_{s=1}^{Kp_n} P(n^{1/2}|\eta_s| > n^{1/2}\lambda_{1n}) \\ &\le O_p(1)\cdot\sum_{s=1}^{Kp_n} \exp\left(-\frac{n\lambda_{1n}^2}{2M_s}\right) \\ &\le O_p(Kp_n)\exp\left(-\frac{n\lambda_{1n}^2}{2\max_s\{M_s\}}\right) \end{aligned}$$

which tends to 0 when $\lambda_{1n} = (2\max_s\{M_s\}\log p_n/n)^{1/2}$. Consequently, by considering the event $\mathcal{F}$, we have

$$|I_1'| \leq \sum_{s=1}^{Kp_n} |\eta_s||\Delta_\mu^{(s)}| \leq \lambda_{1n}\|\Delta_\mu\|_1$$

with a probability tending to one. Note that $|I_3'| \leq \lambda_{1n}\sum_{j=1}^{p_n}\sum_{1\leq k<k'\leq K}|\Delta_{\mu_{kj}} - \Delta_{\mu_{k'j}}| \leq (K-1)\lambda_{1n}\|\Delta_\mu\|_1$. Thus, with a probability tending to one, we have

$$\begin{aligned} |I_1'| + |I_3'| &\leq K\lambda_{1n}\|\Delta_\mu\|_1 \\ &\leq K^{3/2}p_n^{1/2}\lambda_{1n}\|\Delta_\mu\|_2 \\ &= O_p(p_n\log p_n/n). \end{aligned}$$

The proof can be concluded from proving that $I_2' \geq C_2 p_n\log p_n/n$ for some constant C_2.

Since $\|\Omega - \Omega^*\| = o_p(1)$, we have

$$\begin{aligned} I_2' &= (2n)^{-1}\sum_{i=1}^{n}\Delta_\mu^\tau E_i^\tau\Omega^* E_i\Delta_\mu^\tau + o_p(1) \\ &\geq (2\kappa_2)^{-1}\left\{\sum_{k=1}^{K} n_k\|\Delta_{\mu_k}\|_2^2/n\right\} \\ &\geq (2\kappa_2)^{-1}\min_{1\leq k\leq K}\frac{n_k}{n}\cdot\|\Delta_\mu\|_2^2 \\ &= C_2 p_n\log p_n/n \end{aligned}$$

with a probability tending to one. This finishes the proof. □

Before proving Theorem 2, we first prove the following lemma.

LEMMA 8.1. *Let $\|\cdot\|_{FP} : R^K \to R$ be the fused penalty $\|x\|_{FP} = \sum_{1\leq k<k'\leq K}|x_k - x_{k'}|$. Then, $\|\cdot\|_{FP}$ is convex and, for any $x \in R^K$, the subdifferential $\partial\|x\|_{FP}$*

is the set of all vectors $s \in R^K$ *such that*

$$s_i = \sum_{j \neq i} \text{sgn}(x_i - x_j),$$

for $i = 1, \ldots, K$.

Proof. For each $1 \leq j \leq K-1$, let $H^{(j)}$ be a $(K-j) \times K$ matrix with $H^{(j)}_{ii} = -1$, $H^{(j)}_{i,i+j} = 1$ for $1 \leq i \leq K-j$ and 0 otherwise. Denote by H the $K(K-1)/2 \times K$ matrix with jth row block matrix $H^{(j)}$. Then, for any $x \in R^K$, $\|x\|_{FP} = \|Hx\|_1$. Note that the l_1 norm $\|\cdot\|_1$ is convex and $\|\cdot\|_{FP}$ is the composition of a linear functional by the l_1 norm. Hence, $\|\cdot\|_{FP}$ is convex. Further, by the definition of the subdifferential of the l_1 norm, for any $y \in R^K$,

$$\|Hy\|_1 \leq \|Hx\|_1 + < H(y-x), \upsilon > \tag{8.1}$$

holds if and only if $\upsilon \in \mathcal{W}_\upsilon \subset R^{K(K-1)/2}$, where $\mathcal{W}_\upsilon$ is the set of all vectors $\upsilon = \text{sgn}(Hx)$. Note that

$$\begin{aligned} < H(y-x), \text{sgn}(Hx) > &= \sum_{1 \leq k < k' \leq K} [(y_{k'} - x_{k'}) - (y_k - x_k)]\text{sgn}(x_{k'} - x_k) \\ &= 2^{-1} \sum_{k' \neq k} [(y_{k'} - x_{k'}) - (y_k - x_k)]\text{sgn}(x_{k'} - x_k) \\ &= \sum_{k=1}^{K} (y_k - x_k) \left\{ \sum_{k' \neq k} \text{sgn}(x_k - x_{k'}) \right\}. \end{aligned}$$

Thus, equation (8.1) is equivalent to

$$\|y\|_{FP} \leq \|x\|_{FP} + < y - x, s >,$$

where s is a K-dimensional vector with ith component $s_i = \sum_{j \neq i} \text{sgn}(x_i - x_j)$. The set of all such vectors s is, therefore, $\partial \|x\|_{FP}$. □

Proof of Theorem 2. First, we prove the sparsistency of the precision matrix estimator $\hat{\Omega}$. The derivative of $Q_n(\omega, \mu, \Omega)$ w.r.t. Ω_{jl} for $(j,l) \in \mathcal{A}^c, j \neq l$ at $(\hat{\omega}, \hat{\mu}, \hat{\Omega})$ is

$$\frac{\partial Q_n(\hat{\omega}, \hat{\mu}, \hat{\Omega})}{\partial \Omega_{jl}} = \hat{\sigma}_{jl} - s_{jl} - 2\lambda_{2n}\mathrm{sgn}(\hat{\Omega}_{jl}),$$

where s_{jl} is the (j,l)th element of $S = S(\hat{\mu})$ and $\mathrm{sgn}(a)$ denotes the sign of a. Note that

$$\begin{aligned} S &= S(\mu^*) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\Delta_{\mu_k}(x_i - \mu_k^*)^\tau \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(x_i - \mu_k^*)\Delta_{\mu_k}^\tau + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\Delta_{\mu_k}\Delta_{\mu_k}^\tau \\ &\triangleq I_1 - I_2 - I_3 + I_4. \end{aligned}$$

Then, we decompose $\hat{\sigma}_{jl} - s_{jl} = A_1 + A_2 + A_3$, where

$$A_1 = \hat{\sigma}_{jl} - \sigma_{jl}^*,\ A_2 = \sigma_{jl}^* - I_{1jl},\ A_3 = I_{2jl} + I_{3jl} - I_{4jl},$$

where B_{jl} denotes the (j,l)th element of matrix B. Now, consider the order of A_1. Under condition (A), we have $\|\Sigma^*\| = O(1)$ and $\|\hat{\Sigma}\| \leq (\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*))^{-1} = O_p(1)$. Thus,

$$\begin{aligned} |A_1| &\leq \|\hat{\Sigma} - \Sigma^*\| \\ &\leq \|\hat{\Sigma}\| \cdot \|\hat{\Omega} - \Omega^*\| \cdot \|\Sigma^*\| \\ &= O_p(\rho_{n2}^{1/2}). \end{aligned}$$

By Lemma 3 in Bickel and Levina (2008), we have $|A_2| = O_p(\{\log p_n/n\}^{1/2})$. Now, we estimate the order of A_3. Since $\max_{1\leq j\leq p_n}\|\hat{\mu}_{(j)} - \mu^*{}_{(j)}\|_2^2 = O_p(\rho_{n1})$

for a sequence $\rho_{n1} \to 0$, we have

$$\begin{aligned} |I_{2jl}| &= \left| n^{-1} \sum_{i=1}^{n} z_{il} \left(\sum_{k=1}^{K} \tau_{ik} \Delta_{\mu_{kj}} \right) \right| \\ &\leq O_p(1) \cdot \left(\sum_{k=1}^{K} n_k \Delta_{\mu_{kj}}^2 / n \right)^{1/2} \\ &\leq O_p(1) \cdot \left(\sum_{k=1}^{K} \Delta_{\mu_{kj}}^2 \right)^{1/2} = O_p(\rho_{n1}^{1/2}). \end{aligned}$$

Similarly, we have $|I_{3jl}| \leq O_p(\rho_{n1}^{1/2})$ and $|I_{4jl}| \leq O_p(\rho_{n1})$. Thus, $|A_3| \leq O_p(\rho_{n1}^{1/2})$. Combining above results yields that

$$\max_{j,l} |\hat{\sigma}_{jl} - s_{jl}| = O_p(\{\log p_n/n\}^{1/2} + \rho_{n1}^{1/2} + \rho_{n2}^{1/2}).$$

Hence, we need to have $\log p_n/n + \rho_{n1} + \rho_{n2} = O(\lambda_{2n}^2)$ in order to have the sign of $\partial Q_n(\hat{\omega}, \hat{\mu}, \hat{\Omega})/\partial \Omega_{jl}$ that depends on $\text{sgn}(\hat{\Omega}_{jl})$ with a probability tending to one. This completes the proof of Theorem 2(1).

Next, we prove the second result of Theorem 2. The main idea of the proof is inspired by Rinaldo (2009). Let $\bar{\tau}_k = n^{-1} \sum_{i=1}^{n} \tau_{ik}$, for $1 \leq k \leq K$. Then, by Lemma 8.1, we know that

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} \tau_{ik} x_i - \lambda_{1n} \bar{\tau}_k^{-1} \hat{\Sigma} \hat{s}_k$$

where $\hat{s}_k = (\hat{s}_{k1}, \ldots, \hat{s}_{kp_n})^\tau$ with jth element $\hat{s}_{kj} = \sum_{t \neq k} \text{sgn}(\hat{\mu}_{kj} - \hat{\mu}_{tj})$. Hence, for $1 \leq k < k' \leq K$,

$$\hat{\mu}_{k'j} - \hat{\mu}_{kj} = \sum_{i=1}^{n} \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k} \right) x_{ij} - \lambda_{1n} e_j^\tau \hat{\Sigma} (\bar{\tau}_{k'}^{-1} \hat{s}_{k'} - \bar{\tau}_k^{-1} \hat{s}_k)$$

where e_k is a p_n-dimensional unit vector with the kth component 1. Since $\lambda_{\max}(\hat{\Sigma}) = \|\hat{\Sigma}\| \leq (\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*))^{-1} \leq \kappa_2$ and $|\bar{\tau}_{k'}^{-1}\hat{s}_{k'l} - \bar{\tau}_k^{-1}\hat{s}_{kl}| \leq 2(K-1)$ for $l = 1, \ldots, p_n$, we have

$$\begin{aligned} \|e_j^\tau \hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)\|_2 &\leq \lambda_{\max}(\hat{\Sigma})\|\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k\|_2 \\ &\leq 2p_n^{1/2}\kappa_2(K-1). \end{aligned} \tag{8.2}$$

As a result, the event $\{\hat{\mathcal{B}} = \mathcal{B}\}$ occurs in probability if both

$$\max_{\mathcal{B}} \left| \sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k} \right) x_{ij} \right| < 2\lambda_{1n} p_n^{1/2} \kappa_2 (K-1) \tag{8.3}$$

and

$$\min_{\mathcal{B}^c} \left| \sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k} \right) x_{ij} - \lambda_{1n} e_j^\tau \hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k) \right| > 0 \tag{8.4}$$

hold with a probability tending to 1 and $n \to \infty$.

We first consider (8.3). For the sake of simplicity, let $M = 2\kappa_2(K-1)$ and $a_{kk'i} = \tau_{ik'}/n_{k'} - \tau_{ik}/n_k$ for $1 \leq i \leq n$. Then, by condition (C)(i), we know that

$$\max_{\mathcal{B}} \left| \sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k} \right) x_{ij} \right| \leq \max_{\mathcal{B}} \left| \sum_{i=1}^n a_{kk'i}\epsilon_{ij} \right| + o_p(\lambda_{1n} p_n^{1/2}),$$

where $\epsilon_{ij} = x_{ij} - \sum_{k=1}^K \tau_{ik}\mu_{kj}^*$, which follows normal distribution with mean 0 and variance σ_{jj}^*. Let $\xi_j^{kk'} = \sum_{i=1}^n a_{kk'i}\epsilon_{ij}$, for $1 \leq k < k' \leq K$ and $j = 1, \ldots, p_n$. It is easy to show that $E\xi_j^{kk'} = 0$, $\mathrm{Var}(\xi_j^{kk'}) = \sum_{i=1}^n a_{kk'i}^2 \sigma_{jj}^* \leq 2\sigma_{jj}^*$, and $\mathrm{Cov}(\xi_j^{kk'}, \xi_t^{ll'}) = \sum_{i=1}^n a_{kk'i}a_{ll't}\sigma_{jt}^*$ for each $(k, k', j) \neq (l, l', t)$. For $(k, k', j) \in \mathcal{B}$, let $\zeta_j^{kk'} \sim N(0, \sum_{i=1}^n a_{kk'i}^2 \sigma_{jj}^*)$ such that

$$\begin{aligned} E(\zeta_j^{kk'})^2 &= E(\xi_j^{kk'})^2, \qquad \text{for all } (k, k', j) \in \mathcal{B}, \\ E(\zeta_j^{kk'}\zeta_t^{ll'}) &\geq E(\xi_j^{kk'}\xi_t^{ll'}), \quad \text{for all } (k, k', j), (l, l', t) \in \mathcal{B} \text{ and } j \neq t. \end{aligned}$$

Then, by Slepian's inequality (Ledoux and Talagrand 1991) and Chernoff's bound for standard Gaussian variables, we have

$$\begin{aligned} P(\max_{\mathcal{B}} |\xi_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) &\leq P(\max_{\mathcal{B}} |\zeta_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) \\ &\leq \sum_{\mathcal{B}} P(|\zeta_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) \\ &\leq \sum_{\mathcal{B}} 2 \exp\left\{ -\frac{\lambda_{1n}^2 p_n M^2}{4 b_{\max}^*} \right\} \\ &= 2 \exp\left\{ -\frac{\lambda_{1n}^2 p_n M^2}{4 b_{\max}^*} + \log |\mathcal{B}| \right\}, \end{aligned}$$

which vanishes under condition (C)(i).

In order to verify (8.4), it is sufficient to show that

$$\max_{\mathcal{B}^c} \left| \sum_{i=1}^n a_{kk'i} \epsilon_{ij} - \lambda_{1n} e_j^\tau \hat{\Sigma} (\bar{\tau}_{k'}^{-1} \hat{s}_{k'} - \bar{\tau}_k^{-1} \hat{s}_k) \right| \leq \alpha_n^{\min},$$

with probability tending to one as $n \to \infty$. Using the triangle inequality, we only need to show that

$$\max_{\mathcal{B}^c} \left| \lambda_{1n} e_j^\tau \hat{\Sigma} (\bar{\tau}_{k'}^{-1} \hat{s}_{k'} - \bar{\tau}_k^{-1} \hat{s}_k) \right| \leq \alpha_n^{\min}/2 \tag{8.5}$$

and

$$\max_{\mathcal{B}^c} \left| \sum_{i=1}^n a_{kk'i} \epsilon_{ij} \right| \leq \alpha_n^{\min}/2. \tag{8.6}$$

Because of (8.2), it is easy to see that the inequality (8.5) holds under condition (C)(ii). Then, we turn to (8.6). For $(k, k', j) \in \mathcal{B}^c$, let $\zeta_j^{kk'} \sim N(0, 2b_{\max}^*)$ so that

$$\begin{aligned} E(\zeta_j^{kk'})^2 &= E(\xi_j^{kk'})^2, \qquad \text{for all } (k, k', j) \in \mathcal{B}^c, \\ E(\zeta_j^{kk'} \zeta_t^{ll'}) &\geq E(\xi_j^{kk'} \xi_t^{ll'}), \quad \text{for all } (k, k', j), (l, l', t) \in \mathcal{B}^c \text{ and } j \neq t. \end{aligned}$$

Then, again, by Slepian's inequality and Chernoff's bound for standard Gaussian variables, we have

$$\begin{aligned} P(\max_{\mathcal{B}^c} |\xi_j^{kk'}| \geq \alpha_n^{\min}/2) &\leq P(\max_{\mathcal{B}^c} |\zeta_j^{kk'}| \geq \alpha_n^{\min}/2) \\ &\leq \sum_{\mathcal{B}^c} 2\exp\left\{-\frac{(\alpha_n^{\min})^2}{16 b_{\max}^*}\right\} \\ &= 2\exp\left\{-\frac{(\alpha_n^{\min})^2}{16 b_{\max}^*} + \log|\mathcal{B}^c|\right\}, \end{aligned}$$

which vanishes if condition (C)(ii) is satisfied. Hence, the proof of Theorem 2(2) is completed. □

Proof of Theorem 3. Given the estimates $\hat{\omega}$, $\hat{\mu}$ and $\hat{\Omega}$ from (2.3)-(2.4), a new observation x^* is assigned to the kth class if

$$x^{*\tau}\hat{\Omega}(\hat{\mu}_k - \hat{\mu}_l) > \log(\hat{\omega}_l/\hat{\omega}_k) + \{(\tilde{\mu}_k + \tilde{\mu}_l)/2\}^\tau \hat{\Omega}(\hat{\mu}_k - \hat{\mu}_l) \tag{8.7}$$

for $l = 1, \ldots, K$ and $l \neq k$, where $\tilde{\mu}_s = \sum_{i=1}^n I(y_i = s)x_i / \sum_{i=1}^n I(y_i = s)$, $s = 1, \ldots, K$.

Given data (y_i, x_i) for $i = 1, \ldots, n$, the conditional misclassification rate of CEDA is given by

$$R_n = \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{(-1)^k \hat{\delta}^\tau \hat{\Omega}(\mu_k^* - \tilde{\mu}_k) - \hat{\delta}^\tau \hat{\Omega}\tilde{\delta}/2}{\sqrt{\hat{\delta}^\tau \hat{\Omega}\Sigma^* \hat{\Omega}\hat{\delta}}}\right),$$

where $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2$ and $\tilde{\delta} = \tilde{\mu}_1 - \tilde{\mu}_2$.

(i) Since $\|\hat{\Omega} - \Omega^*\|^2 = O_p(\rho_{n2})$ for a sequence $\rho_{n2} \to 0$, we have

$$\begin{aligned} \|\hat{\Sigma} - \Sigma^*\| &= \|\hat{\Sigma}(\hat{\Omega} - \Omega^*)\Sigma^*\| \\ &\leq \|\hat{\Sigma}\| \cdot \|\hat{\Omega} - \Omega^*\| \cdot \|\Sigma^*\| \\ &\leq \|\hat{\Sigma}\| \cdot O_p(\kappa_2 \rho_{n2}^{1/2}). \end{aligned}$$

Note that $\|\hat{\Sigma}\| \leq (\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*))^{-1} = O_p(1)$. Hence,

$$\|\hat{\Sigma} - \Sigma^*\|^2 = O_p(\rho_{n2}).$$

Consequently,

$$\hat{\delta}^\tau \hat{\Omega}\Sigma^*\hat{\Omega}\hat{\delta} = \hat{\delta}^\tau \hat{\Omega}\hat{\delta}[1 + O_p(\rho_{n2}^{1/2})] = \hat{\delta}^\tau \Omega^* \hat{\delta}[1 + O_p(\rho_{n2}^{1/2})].$$

Without loss of generality, we assume that $\hat{\delta} = (\hat{\delta}_1^\tau, 0^\tau)^\tau$, where $\hat{\delta}_1$ is the $\hat{b}_n$-dimensional vector containing nonzero components of $\hat{\delta}$. Let $\delta_\mu^* = (\delta_1^{*\tau}, 0^\tau)^\tau$, where δ_1^* is the b_n-dimensional vector containing nonzero components of δ_μ^*. Then, from Theorem 2, we have $\hat{b}_n = b_n$ and consequently,

$$\|\hat{\delta} - \delta_\mu^*\|_2^2 = \|\hat{\delta}_1 - \delta_1^*\|_2^2 = O_p(b_n \rho_{n1})$$

with a probability tending to one. It together with condition (A) implies that $(\hat{\delta} - \delta_\mu^*)^\tau \Omega^* (\hat{\delta} - \delta_\mu^*) = O_p(b_n \rho_{n1})$. Thus, $(\hat{\delta} - \delta_\mu^*)^\tau \Omega^* \delta_\mu^* \leq \Delta_{p_n} O_p(b_n^{1/2} \rho_{n1}^{1/2})$ and

$$\begin{aligned} \hat{\delta}^\tau \Omega^* \hat{\delta} &= (\hat{\delta} - \delta_\mu^*)^\tau \Omega^* (\hat{\delta} - \delta_\mu^*) + 2(\hat{\delta} - \delta_\mu^*)^\tau \Omega^* \delta_\mu^* + \Delta_{p_n}^2 \\ &= \Delta_{p_n}^2 [1 + O_p(b_n^{1/2} \rho_{n1}^{1/2} / \Delta_{p_n})]. \end{aligned}$$

Let $\tilde{\mu}_1 - \mu_1^* = (\gamma_1^\tau, \gamma_2^\tau)^\tau$, where γ_1 is a b_n-dimensional vector. Partition Ω^* into

$$\Omega^* = \begin{bmatrix} \Omega_{11}^* & \Omega_{12}^* \\ \Omega_{12}^{*\tau} & \Omega_{22}^* \end{bmatrix},$$

where Ω_{11}^* is a $b_n \times b_n$ matrix, and partition Σ^*, $\hat{\Omega}$ and $\hat{\Sigma}$ in the same way. Then,

$$\hat{\delta}^\tau \hat{\Omega}(\tilde{\mu}_1 - \mu_1^*) = \hat{\delta}_1^\tau \hat{\Omega}_{11}\gamma_1 + \hat{\delta}_1^\tau \hat{\Omega}_{12}\gamma_2,$$

with a probability tending to one. Further, by Cauchy-Schwarz inequality and the fact $\Omega_{11}^{*-1} \leq \Sigma_{11}^*$, we have $(\hat{\delta}_1^\tau \hat{\Omega}_{11}\gamma_1)^2 \leq (\hat{\delta}^\tau\hat{\Omega}\hat{\delta})O_p(b_n/n)$ and $(\hat{\delta}_1^\tau \hat{\Omega}_{12}\gamma_2)^2 \leq (\hat{\delta}^\tau\hat{\Omega}\hat{\delta})\{\gamma_2^\tau\Omega_{12}^{*\tau}\Sigma_{11}^*\Omega_{12}^*\gamma_2[1+O_p(\rho_{n2}^{1/2})]\}$. Note that all eigenvalues of sub-matrices of Ω^* and Σ^* are bounded under condition (A). Then, we have that

$$\begin{aligned} E(\gamma_2^\tau\Omega_{12}^{*\tau}\Sigma_{11}^*\Omega_{12}^*\gamma_2) &\leq \kappa_2 E(\gamma_2^\tau\Omega_{12}^{*\tau}\Omega_{12}^*\gamma_2) \\ &\leq \frac{\kappa_2^2}{n}\mathrm{tr}(\Omega_{12}^*\Omega_{12}^{*\tau}) \\ &\leq \kappa_2^2 a_n/n. \end{aligned}$$

Therefore,

$$\frac{\hat{\delta}^\tau\hat{\Omega}(\tilde{\mu}_1-\mu_1^*)}{\sqrt{\hat{\delta}^\tau\hat{\Omega}\Sigma^*\hat{\Omega}\hat{\delta}}} = \frac{O_p(\sqrt{b_n/n})+O_p(\sqrt{a_n/n})}{\sqrt{1+O_p(\rho_{n2}^{1/2})}},$$

which also holds when $\tilde{\mu}_1-\mu_1^*$ is replaced by $\tilde{\mu}_2-\mu_2^*$ or $\tilde{\delta}-\delta_\mu^*$. Furthermore, $\hat{\delta}^\tau\hat{\Omega}\tilde{\delta} = \hat{\delta}^\tau\hat{\Omega}\hat{\delta}+\hat{\delta}^\tau\hat{\Omega}(\tilde{\delta}-\delta_\mu^*)+\hat{\delta}^\tau\hat{\Omega}(\delta_\mu^*-\hat{\delta})$ and $[\hat{\delta}^\tau\hat{\Omega}(\delta_\mu^*-\hat{\delta})]^2 \leq (\hat{\delta}^\tau\Omega^*\hat{\delta})O_p(b_n\rho_{n1})$. Therefore,

$$\begin{aligned} \frac{(-1)^k\hat{\delta}^\tau\hat{\Omega}(\mu_k^*-\tilde{\mu}_k)-\hat{\delta}^\tau\hat{\Omega}\tilde{\delta}/2}{\sqrt{\hat{\delta}^\tau\hat{\Omega}\Sigma^*\hat{\Omega}\hat{\delta}}} &= \frac{O_p(\sqrt{b_n/n})+O_p(\sqrt{a_n/n})+O_P(\sqrt{b_n\rho_{n1}})}{\sqrt{1+O_p(\rho_{n2}^{1/2})}} \\ &\quad -\frac{\Delta_{p_n}\sqrt{1+O_p(b_n^{1/2}\rho_{n1}^{1/2}/\Delta_{p_n})}}{2\sqrt{1+O_p(\rho_{n2}^{1/2})}} \\ &= -[1+O_p(c_n)]\Delta_{p_n}/2, \end{aligned}$$

which implies the result in (i).

(ii) Let ϕ be the density of Φ. Then, by the result in (i),

$$R_n - R_{\mathrm{OPT}} = \phi(\nu_n)O_p(c_n),$$

where ν_n is between $-\Delta_{p_n}/2$ and $-[1+O_p(c_n)]\Delta_{p_n}/2$. Since Δ_{p_n} is bounded, $\phi(\nu_n)$ is bounded by a constant and R_{OPT} is bounded away from 0. Hence, the CEDA is asymptotically optimal and $R_n/R_{\mathrm{OPT}} - 1 = O_p(c_n)$.

(iii) When $\Delta_{p_n} \to \infty$, $R_{\mathrm{OPT}} \to 0$ and by the result in (i), $R_n \xrightarrow{P} 0$. Thus, the CEDA is asymptotically sub-optimal.

(iv) If $\Delta_{p_n} \to \infty$ and $c_n\Delta_{p_n}^2 \to 0$, then, by Lemma 1 in Shao et al. (2011), we have $R_n/R_{\mathrm{OPT}} \xrightarrow{P} 1$. □